On-Line Parameter Estimation in General State-Space Models

Christophe Andrieu
School of Mathematics
University of Bristol, UK.
c.andrieu@bris.ac.uk

Arnaud Doucet
Dpts of CS and Statistics
Univ. of British Columbia, CA.
arnaud@stat.ubc.ca

Vladislav B. Tadić
Dpt of Automatic Control,
University of Sheffield, UK.
v.tadic@sheffield.ac.uk

***Abstract*— The estimation of static parameters in general non-linear non-Gaussian state-space models is a long-standing problem. This is despite the advent of Sequential Monte Carlo (SMC, aka particle filters) methods, which provide very good approximations to the optimal filter under weak assumptions. Several algorithms based on SMC have been proposed in the past 10 years to solve the static parameter problem. However all the algorithms we are aware of suffer from the so-called 'degeneracy problem'. We propose here a methodology for point estimation of static parameters which does not suffer from this problem. Our methods take advantage of the fact that many state space models of interest are ergodic and stationary: this allows us to propose contrast functions for the static parameter which can be consistently estimated and optimised using simulation-based methods. Several types of contrast functions are possible but we focus here on the average of a so-called pseudo-likelihood which we maximize using on-line Expectation-Maximization type algorithms. In its basic form the algorithm requires the expression of the invariant distribution of the underlying state process. When the invariant distribution is unknown, we present an alternative which relies on indirect inference techniques.**

I. Introduction

This paper is concerned with the on-line estimation of static parameters in non-linear non-Gaussian state-space models. More precisely, we consider models of the following form. For any parameter $\theta \in \Theta$, the hidden/latent state process $\{X_n; n \geq 1\} \subset \mathsf{X}^{\mathbb{N}}$ is a *stationary and ergodic* Markov process, characterized by its Markov transition probability distribution $f_\theta(x'|\,x)$ admitting π_θ as invariant distribution, *i.e.* $X_1 \sim \pi_\theta$ and for $n \geq 1$,

$$X_{n+1}|\,(X_n = x) \sim f_\theta(\cdot|\,x) \;\; . \tag{1}$$

Note that the assumption $X_1 \sim \pi_\theta$ is not restrictive under the ergodicity assumption, and will furthermore simplify the presentation of our methodology. As indicated by its name $\{X_n\}$ is observed, not directly, but through another process $\{Y_n; n \geq 1\} \subset \mathsf{Y}^{\mathbb{N}}$. The observations are assumed to be conditionally independent given $\{X_n\}$, and their common marginal probability distribution is of the form $g_\theta(y|\,x)$; *i.e.* for $1 \leq n \leq m$,

$$Y_n|\,(X_1, \ldots, X_n = x, \ldots, X_m) \sim g_\theta(\cdot|\,x) \;\; . \tag{2}$$

We give here a couple of standard examples used throughout this paper.

Example 1. *Linear Gaussian model*

$$X_{n+1} = \phi X_n + \sigma_v V_{n+1},\; V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \;\; ,$$
$$Y_n = X_n + \sigma_w W_n,\; W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \;\; ,$$

where $\Theta = (-1,1) \times \mathbb{R}^+ \times \mathbb{R}^+$, $\theta = (\phi, \sigma_v^2, \sigma_w^2)$ denotes the static parameter vector and $\mathcal{N}(x; \mu, \sigma^2)$ is the normal distribution with argument x, mean μ and variance σ^2. It can easily be checked that $\pi_\theta(x) = \mathcal{N}\left(x; \frac{\sigma_v^2}{1-\phi^2}\right)$, $f_\theta(x'|x) = \mathcal{N}(x'; \phi x, \sigma_v^2)$ and $g_\theta(y|x) = \mathcal{N}(y; x, \sigma_w^2)$.

Example 2. *Stochastic volatility model* [10]

$$X_{n+1} = \phi X_n + \sigma_v V_{n+1}, Y_n = \beta \exp(X_n/2) W_n,$$

where $\Theta = (-1,1) \times \mathbb{R}^+ \times \mathbb{R}^+$ and $\theta = (\phi, \sigma_v^2, \beta^2)$ denotes the static parameter vector. The transition probability $f_\theta(x'|x)$ and invariant distribution $\pi_\theta(x)$ are identical to those of the previous example but here $g_\theta(y|x) = \mathcal{N}(y; 0, \beta^2 \exp(x))$.

When the static parameter θ is known, sequential inference on the process $\{X_n\}$ is typically based on the sequence of joint posterior distributions $\{p_\theta(x_{1:n}|Y_{1:n})\}$ which each summarizes all the information collected about $X_{1:n}$ up to time n. *Optimal filtering* is concerned with the sequential estimation of these distributions, which can be - at least conceptually - easily achieved using the following updating formula for $n \geq 2$

$$p_\theta(x_{1:n}|Y_{1:n}) = \frac{g_\theta(Y_n|x_n) f_\theta(x_n|x_{n-1})}{p_\theta(Y_n|Y_{1:n-1})} p_\theta(x_{1:n-1}|Y_{1:n-1}) \;\; , \tag{3}$$

and $p_\theta(x_1|Y_1) \propto g_\theta(Y_1|x_1) \pi_\theta(x_1)$. Although simple, the recursion formula in Eq. (3) rarely admits a closed form expression: this is typically the case as soon as f_θ or g_θ are non-Gaussian, or X is not a finite set. In such scenarios it is possible to resort to numerical approximations. One such class of numerical algorithms are Sequential Monte Carlo (SMC) methods (aka particle filters), which have recently proved to be efficient tools to propagate sample approximations of these distributions' marginals $\{p_\theta(x_{n-L+1:n}|Y_{1:n})\}$ in time for a given integer $L > 0$ [5]. This methodology is now well developed and the theory supporting this approach is also well established [4].

We focus in this paper on the on-line estimation of the static parameter θ. More precisely, assuming that there is a true parameter value θ^* generating the data $\{Y_n\}$, and that this value is unknown, our aim is to compute point estimates of θ^* from $\{Y_n\}$ in an on-line manner. This problem appears in numerous applications. First, in most real-world scenarios θ^* is indeed unknown and its estimation is required before optimal filtering can be performed. Second, on-line estimation is often the only realistic solution when the amount of data to be processed is large. Although apparently simpler than the optimal filtering problem, the static parameter estimation problem has proved to be much more difficult; no closed form solutions are, in general, available, even for linear Gaussian and finite state-space hidden Markov models. There have already been numerous attempts to solve it in control, signal processing, statistics and related fields; *e.g.* [6], [7], [13]. However it remains largely unsolved despite the possibilities offered by SMC techniques. We propose here a general and principled methodology which allows us to compute asymptotically consistent point estimates of θ^* for a large class of dynamic models. Our approach is essentially based upon the on-line maximization of a pseudo-likelihood function for which Monte Carlo simulations might be needed. However, we would like to stress at this point that the methodology developed here does not necessarily require the use of such computationally intensive approaches when more direct and simpler simulation techniques are possible.

II. SMC Methods for Static Parameter Estimation

It is not our aim here to review SMC methods in details, but simply to point out their intrinsic limitations which have fundamental practical consequences for the static parameter estimation problem. Assuming that the static parameter θ is fixed for the time being, we describe the simplest SMC algorithm available to approximate $\{p_\theta(x_{1:n}|Y_{1:n})\}$ sequentially. More elaborate algorithms are reviewed in [5], but crucially all such SMC algorithms suffer from a common problem, namely *path degeneracy*, as explained below.

A. Sampling Importance Resampling

Assume that at time $n-1$, a collection of N ($N \gg 1$) random samples $\{\hat{X}_{1:n-1}^{(i)}, i = 1, \ldots, N\}$, called particles, distributed approximately according to $p_\theta(x_{1:n-1}|Y_{1:n-1})$ is available. The empirical distribution

$$\hat{p}_\theta^N(dx_{1:n-1}|Y_{1:n-1}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{X}_{1:n-1}^{(i)}}(dx_{1:n-1}) \tag{4}$$

is an approximation of $p_\theta(dx_{1:n-1}|Y_{1:n-1})$, where $\delta_{x_0}(dx)$ represents the Dirac delta mass function located in x_0. Now at time n, we wish to produce N particles which will define an approximation $\hat{p}_\theta^N(dx_{1:n}|Y_{1:n})$ of $p_\theta(dx_{1:n}|Y_{1:n})$. A simple method to achieve this consists of setting $\tilde{X}_{1:n-1}^{(i)} = \hat{X}_{1:n-1}^{(i)}$ and then sample, for example, $\tilde{X}_n^{(i)} \sim f_\theta(\cdot|\tilde{X}_{n-1}^{(i)})$. The resulting empirical distribution of the particles $\{\tilde{X}_{1:n}^{(i)}\}$ is an approximation of the joint density $p(x_{1:n-1}|Y_{1:n-1}) f_\theta(x_n|x_{n-1})$. We correct for the discrepancy between this density and the target $p_\theta(x_{1:n}|Y_{1:n})$ using importance sampling. This yields the following approximation of $p(x_{1:n}|Y_{1:n})$

$$\tilde{p}_\theta^N(dx_{1:n}|Y_{1:n}) = \sum_{i=1}^{N} W_n^{(i)} \delta_{\tilde{X}_{1:n}^{(i)}}(dx_{1:n}) \ , \tag{5}$$

where $W_n^{(i)} \propto g_\theta(Y_n|\tilde{X}_n^{(i)})$ and $\sum_{i=1}^{N} W_n^{(i)} = 1$.

To obtain an unweighted approximation of $p(x_{1:n}|Y_{1:n})$ of the form (4), we resample particles $\{\tilde{X}_{1:n}^{(i)}\}$ according to probabilities proportional to their weights $\{W_n^{(i)}\}$. The underlying idea is to get rid of particles with small weights and multiply particles which are in the regions with high probability masses; see [5].

B. Limitations of SMC Methods

Under relatively weak assumptions on f_θ and g_θ, it can be proved that the resulting set of empirical posterior distributions $\{\hat{p}_\theta^N(dx_{1:n}|Y_{1:n})\}$ converge towards the true posteriors as N goes to infinity. More precisely, it can easily be shown that for any $n \geq 1$ and any bounded test function $\varphi_n : \mathsf{X}^n \to \mathbb{R}$ there exists some constant $C_{\theta,n}(\varphi_n) < \infty$ such that for any $N \geq 1$

$$\mathbb{E}\left[\left(\int_{\mathsf{X}^n} \varphi_n(x_{1:n})\, \epsilon_\theta^N(dx_{1:n}|Y_{1:n})\right)^2\right] \leq \frac{C_{\theta,n}(\varphi_n)}{N} \ , \tag{6}$$

where $\epsilon_\theta^N := p_\theta - \widehat{p}_\theta^N$ and the expectation is with respect to the particle realizations, see [4]. Although at first sight reassuring, (6) is practically useless since the bound $C_{\theta,n}(\varphi_n)$ typically grows polynomially or exponentially with n, and reflects a fundamental weakness of SMC methods: with limited resources, *i.e.* N fixed and finite, it is not possible to approximate properly the sequence of distributions $\{p_\theta(x_{1:n}|Y_{1:n})\}$.

We now illustrate, with a toy example, the underlying phenomenon which explains the growth of $\{C_{\theta,n}(\varphi_n)\}$. The tree in Fig. 1 represents a realization of the paths $\{\hat{X}_{1:n}^{(i)}\}$ of $N = 8$ particles up to time $n = 8$ for a system for which the state space is $\mathsf{X} = \{-5, -4, \ldots, 0, 1, \ldots, 5\}$. The numbers at each node represent the number of particles that effectively pass through it. This realization of the particle process is representative of what is generally observed in more complex scenarios: the paths tend to coalesce as we follow the paths backward in time. As a result, whereas $\{\hat{X}_8^{(i)}\}$ and $\{\hat{X}_7^{(i)}\}$ provide a good coverage of X, which will result in a good representation of $p(x_8|Y_{1:8})$ and $p(x_7|Y_{1:8})$, the sample representation deteriorates as we go back in time, resulting in poor sample approximations of $p(x_{1:4}|Y_{1:8})$, *i.e.* even if the true $p(x_{1:4}|Y_{1:8})$ is not degenerate, the sample representation is degenerate. This coalescence phenomenon is the result of the resampling stage and has long been observed. As we shall see, this inability of SMC methods to satisfactorily approximate (*i.e.* with a constant computational budget per iteration) the sequence of joint distributions $\{p_\theta(x_{1:n}|Y_{1:n})\}$ makes SMC-based on-line parameter estimation algorithms inappropriate.

The success of SMC methods lies in the fact that results of the following form can be obtained under the relatively general assumptions detailed in [4]. Let $L > 0$ be an integer and let $\varphi_L : \mathsf{X}^L \to \mathbb{R}$ be a bounded test function, then there exists some constant $D_{\theta,L}(\varphi_L) < \infty$ such that for any $n \geq 1$,

$$\mathbb{E}\left[\left(\int_{\mathsf{X}^L} \varphi_L(x_{n-L+1:n})\, \epsilon_\theta^N(dx_{n-L+1:n}|Y_{1:n})\right)^2\right] \leq \frac{D_{\theta,L}(\varphi_L)}{N} \ . \tag{7}$$

In summary, for a fixed computational budget per time instant, SMC methods cannot properly approximate joint distributions sequences of the form $\{p_\theta(x_{1:n}|Y_{1:n})\}$ sequentially in time because of the paths' coalescence phenomenon: as we shall see this is what makes the direct application of SMC techniques inappropriate for static parameter inference. However, under ergodic assumptions, for a given lag $L > 0$, SMC methods can consistently approximate sequences of distributions $\{p_\theta(x_{n-L+1:n}|Y_{1:n})\}$ for a fixed number N of particles: this is the type of property that we shall use in this paper in order to perform consistent on-line static parameter estimation.

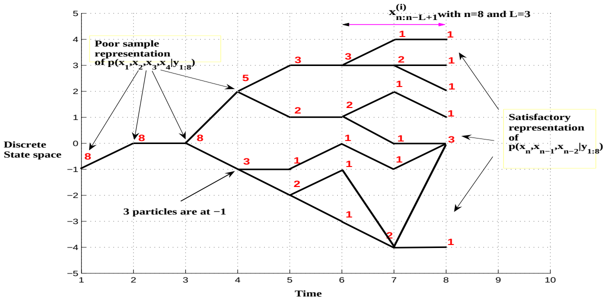
$x^{(i)}_{n:n-L+1}$ with n=8 and L=3
Poor sample representation of $p(x_1,x_2,x_3,x_4|y_{1:8})$
Satisfactory representation of $p(x_n,x_{n-1},x_{n-2}|y_{1:8})$
Discrete State space
3 particles are at −1
Time
8
8
8
5
3
3
3
2
2
2
1
1
1
1
1
1
1
1
1
1
1
1
1
3
1
5
4
3
2
1
0
−1
−2
−3
−4
−5
1
2
3
4
5
6
7
8
9
10

Fig. 1. Realistic sequential methods suffer from path degeneracy.

C. Difficulties with Static Parameters

Various strategies have been proposed in order to deal with static parameter in an SMC context. These methods are reviewer in the extended version of this paper, and we focus here on a popular method where one sets a prior on θ and SMC is used to estimate the joint posterior $p(\theta, x_{1:n}|Y_{1:n})$. Diversity among particles in the parameter space is introduced using MCMC steps of invariant distributions $p(\theta|Y_{1:n}, x_{1:n})$. This is certainly one of the most elegant method, as the model of interest is not artificially altered. This algorithm takes a simple form when $p(Y_{1:n}|x_{1:n}, \theta)$ can be summarized by a set of low-dimensional sufficient statistics [1], [6], [7]. However, as noticed in [1] and in light of the limitations of SMC methods outlined previously, this approach is inefficient. This is demonstrated by the following example.

Example 1 *Linear Gaussian model* (cont.). For the sake of simplicity, we set a uniform prior distribution on the stability domain $(-1,1)$ for ϕ and we assume that $X_1 \sim \mathcal{N}\left(0,\sigma_0^2\right)$ for some $\sigma_0^2 > 0$. In this case, the full conditional distribution $p\left(\phi\right| Y_{1:n}, x_{1:n})$ is a truncated Gaussian distribution restricted to $(-1,1)$ with mean m_n and variance σ_n^2 given by

$$m_n = \sigma_n^2 \left(\sum_{k=2}^{n} x_k x_{k-1}\right) \text{ and } \sigma_n^{-2} = \sum_{k=1}^{n-1} x_k^2 ,$$

The problem with this approach is that the SMC estimates of the sufficient statistics necessary to perform the MCMC updates degrade as n increases because they are based on the approximation of the joint distribution $p_{\phi_*}\left(x_{1:n}\right| Y_{1:n})$. For the ideal case where $\phi = \phi^* = 0.5$ we present in Figure 2 the quantity $n^{-1}\sum_{k=1}^{n}\mathbb{E}\left[X_k^2\right| Y_{1:n}]$ computed exactly using the Kalman smoother and estimated using an SMC method. Initially the SMC estimate displays good performance but, as expected, performs very poorly as n increases: this stems from the fact that the joint distributions $\{p_{\phi^*}\left(x_{1:n}\right| Y_{1:n})\}$ cannot be consistently estimated over time, as pointed out earlier. In Figure 2, we display the parameter estimate obtained using the SMC algorithm coupled with Gibbs sampling updates. We see that at first the parameter seems to converge towards the correct region but then drifts away as the sufficient statistics used in the MCMC update are not properly estimated. A similar problem occurs with the method proposed in [13] since it is also based on such sufficient statistics.

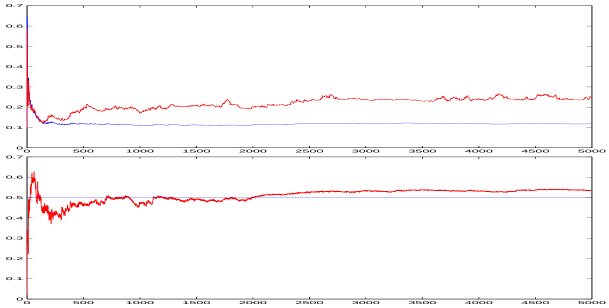

Fig. 2. Top: sufficient statistics computed exactly through the Kalman filter (solid line) and the SMC algorithm (dotted line). Bottom: parameter estimated using SMC methods combined with Gibbs steps

III. Point Estimation Methods

We present here an alternative strategy to the static parameter estimation problem, which aims to give point estimates of θ^* rather than a series of estimates of the posterior distributions $\{p(\theta|Y_{1:n})\}$. As a result no particle method is required in the parameter space, and it should also be pointed out that particle methods in the state space X are also not always necessary. The most natural approach to point estimation for the parameter θ^* consists of recursively maximizing the series of likelihoods $\{p(Y_{1:n}|\theta)\}$. We start this section with a discussion in which we highlight the difficulties associated with this type of strategy and this leads us to instead focus on a pseudo-likelihood approach which is well suited to Monte Carlo approximations. We then go on to develop efficient algorithms for maximizing the pseudo likelihood recursively. We first describe a gradient algorithm in brief, and then focus mainly on on-line EM (Expectation-Maximization) type algorithms which benefit from numerical stability.

A. Likelihood and Pseudo-Likelihood Functions

The log likelihood function corresponding to model (1)-(2) is given, for $n \geq 1$ observations by $\log\, p_\theta\left(Y_{1:n}\right) = \sum_{k=1}^{n}\log\, p_\theta\left(Y_k\right| Y_{1:k-1})$ with the convention $Y_{1:0} = \varnothing$ and where

$$p_\theta\left(Y_k\right| Y_{1:k-1}) = \int_{\mathsf{X}} g_\theta\left(Y_k\right| x_k)\, p_\theta\left(x_k\right| Y_{1:k-1})\, dx_k \ .$$

Under regularity assumptions, it can be shown that the average log-likelihood satisfies the following ergodicity property

$$\begin{array}{l} \lim_{n\rightarrow\infty}\ n^{-1}\log\ p_\theta\left(Y_{1:n}\right) = l\left(\theta\right)\ , \\ l\left(\theta\right) := \int_{\mathsf{Y}\times\mathcal{P}(\mathsf{X})}\log\left(\int_{\mathsf{X}} g_\theta\left(y\right| x)\,\mu\left(dx\right)\right)\lambda_{\theta,\theta^*}\left(dy, d\mu\right)\ , \end{array}$$

where $\mathcal{P}\left(\mathsf{X}\right)$ is the set of probability distributions defined on X. In this expression $\lambda_{\theta,\theta^*}$ is the invariant distribution of the Markov chain $\{Y_k, p_\theta\left(x_k\right| Y_{1:k-1})\}$. It can be shown that the set of global maxima of this function includes the true value θ^*. This follows from the fact that maximizing $l\left(\theta\right)$ is equivalent to minimizing the Kullback-Leibler divergence $K\left(\theta,\theta^*\right) := l\left(\theta^*\right) - l\left(\theta\right) \geq 0$. Based upon this remark, we suggest the use of noisy gradient algorithms to maximize $l\left(\theta\right)$; see [2] for a review. There are, however, two major problems with this approach. First, such an approach requires one to estimate the derivative of the optimal filter with respect to θ. Non-standard particle methods are required to estimate this signed measure and their robust implementation via SMC has a computational complexity in $O\left(N^2\right)$, where N is the number of samples used for the SMC [11]. Second it can be difficult to properly scale the gradient components. More elegant and robust algorithms have been proposed that rely on on-line versions of the EM algorithm, typically when the joint distribution $p_\theta\left(x_{1:n}, y_{1:n}\right)$ belongs to the exponential family. This approach has the advantage that the filter derivative is not required and that it is, in general, numerically well-behaved. Furthermore, it is conceptually and practically straightforward to implement an on-line EM algorithm to maximize $l\left(\theta\right)$ using SMC methods. However, this requires estimating sufficient statistics based on joint probability distributions whose dimension is increasing over time. So similarly to SMC approaches that use the MCMC steps described earlier, such an approach can lead to unsatisfactory results in practice.

To circumvent this problem we propose to introduce here another contrast function. More precisely, we focus on a pseudo-likelihood function as originally proposed in [12] for finite state space HMMs (refered to in this paper as “split-data likelihood”). This pseudo-likelihood is defined as follows. Formally, consider for a given time lag $L \geq 1$ and any $k \geq 0$ “slices” $\mathsf{X}_k = X_{kL+1:(k+1)L}$ and $\mathsf{Y}_k = Y_{kL+1:(k+1)L}$ of $\{X_n\}$ and $\{Y_n\}$. Because of our stationarity assumption, the vectors $\{\mathsf{X}_k, \mathsf{Y}_k\}$ are *identically distributed* and their common distribution $p_\theta\left(\mathsf{x}_k, \mathsf{y}_k\right)$ is given by

$$\pi_\theta(x_{kL+1})g_\theta(y_{kL+1}|x_{kL+1})\prod_{n=kL+2}^{(k+1)L} f_\theta(x_n|x_{n-1})g_\theta(y_n|x_n)\ . \tag{8}$$

The likelihood of a block Y_k of observations is given by

$$p_\theta\left(\mathsf{Y}_k\right) = \int_{\mathsf{X}^L} p_\theta\left(\mathsf{x}_k, \mathsf{Y}_k\right) d\mathsf{x}_k\ , \tag{9}$$

and we define the log pseudo-likelihood for m slices of observations by $\sum\limits_{k=0}^{m-1}\log\ p_\theta\left(\mathsf{Y}_k\right)$ which compared to the true likelihood essentially ignores the dependence between data slices. The parameter L should be large enough to ensure identifiability. Note also that there will be here an efficiency/computational complexity trade-off associated to L. As L increases, the maximum pseudo-likelihood estimate properties will become comparable to that of the standard ML estimate, but as we shall see this might result in more complex and computationally intensive algorithms.

Under ergodicity assumptions, the average log pseudo–likelihood satisfies

$$\begin{array}{l}\lim_{m\to\infty}\frac{1}{m}\sum_{k=0}^{m-1}\log\ p_{\theta}\left(\mathsf{Y}_k\right)=\bar{l}\left(\theta\right)\ \ ,\\ \bar{l}\left(\theta\right):=\int_{\mathsf{Y}^L}\log\left(p_{\theta}\left(\mathsf{y}\right)\right)p_{\theta^*}(\mathsf{y})d\mathsf{y}\ .\end{array}$$

It can be shown that the set of parameters maximizing $\bar{l}\left(\theta\right)$ includes the true parameter [12]. This follows from the fact that maximizing $\bar{l}\left(\theta\right)$ is equivalent to minimizing the following Kullback-Leibler divergence $\bar{K}\left(\theta,\theta^*\right)=\bar{l}\left(\theta^*\right)-\bar{l}\left(\theta\right)\geq 0$. In this article, we propose to maximize $\bar{l}\left(\theta\right)$ recursively using on-line EM type algorithms and stochastic approximation techniques.

Whereas the maximization of the true average log-likelihood function requires complex SMC methods in order to either evaluate the filter derivative or estimate expectations with respect to distributions defined on X^n at time n, the key advantage (detailed further) of the average log pseudo-likelihood function is that it only requires the estimation of expectations with respect to distributions defined on X^L. A direct stochastic steepest descent algorithm to minimize $\bar{K}\left(\theta,\theta^*\right)$ is possible using Fisher's identity. We will not detail this approach here, but will focus on on-line EM type algorithms as they are more numerically stable and widely applicable to models used in practice.

B. *On-line EM Algorithm: Known Invariant Distribution*

We first assume in this section that the invariant distribution π_θ is known analytically. To introduce the on-line EM in this scenario, we first present an "ideal" batch EM algorithm to minimize $\bar{K}\left(\theta,\theta^*\right)$ with respect to θ or equivalently to maximize $\bar{l}\left(\theta\right)$. At iteration $k+1$, given an estimate θ_k of θ^*, we update our estimate via

$$\theta_{k+1}=\underset{\theta\in\Theta}{\arg\max}\ Q\left(\theta,\theta_k\right)\ \ ,$$

where

$$Q\left(\theta,\theta_k\right)=\int_{\mathsf{X}^L\times\mathsf{Y}^L}\log\left(p_\theta\left(\mathsf{x},\mathsf{y}\right)\right)p_{\theta_k}\left(\mathsf{x}|\mathsf{y}\right)p_{\theta^*}(\mathsf{y})d\mathsf{x}d\mathsf{y}\ .$$

Now since for any $\theta\in\Theta$ $Q\left(\theta_{k+1},\theta_k\right)-Q\left(\theta_k,\theta_k\right)-\bar{K}\left(\theta_k,\theta^*\right)+\bar{K}\left(\theta_{k+1},\theta^*\right)$ is equal to

$$\int_{\mathsf{X}^L\times\mathsf{Y}^L}\log\left(\frac{p_{\theta_{k+1}}\left(\mathsf{x}|\mathsf{y}\right)}{p_{\theta_k}\left(\mathsf{x}|\mathsf{y}\right)}\right)p_{\theta_k}\left(\mathsf{x}|\mathsf{y}\right)p_{\theta^*}(\mathsf{y})d\mathsf{x}d\mathsf{y}$$

and since the second term on the RHS is negative by Jensen's inequality, we see that an iteration of this "ideal" EM algorithm decreases the value of $\bar{K}\left(\theta,\theta^*\right)$, and the stationary points correspond to the zeros of $\bar{K}\left(\theta,\theta^*\right)$. In practice for the models which we will consider, it is necessary to compute a set of sufficient statistics $\Phi\left(\theta_k,\theta^*\right)$ at time k in order to compute Q.

Example 2. *Stochastic volatility model* (cont.) In this case, we have (with $\equiv$ meaning "equal up to a constant")

$$\begin{array}{l}Q\left(\theta,\theta_{k-1}\right)\equiv-\log\left(1-\phi^2\right)+L\log\left(\sigma_v^2\beta^2\right)+\frac{1}{\beta^2}\varphi_4\left(\theta_{k-1},\theta^*\right)\\ +\frac{1}{\sigma_v^2}\left[\varphi_1\left(\theta_{k-1},\theta^*\right)+\left(1+\phi^2\right)\varphi_2\left(\theta_{k-1},\theta^*\right)-2\phi\varphi_3\left(\theta_{k-1},\theta^*\right)\right]\end{array}$$

where the sufficient statistics $\Phi\left(\theta_{k-1},\theta^*\right)=\left[\varphi_1\left(\theta_{k-1},\theta^*\right),\varphi_2\left(\theta_{k-1},\theta^*\right),\varphi_3\left(\theta_{k-1},\theta^*\right),\varphi_4\left(\theta_{k-1},\theta^*\right)\right]^T$ are given by $\Phi\left(\theta_{k-1},\theta^*\right)=\mathbb{E}_{\theta_{k-1},\theta^*}\left[\Psi\left(\mathsf{X},\mathsf{Y}\right)\right]$ and where the expectation is with respect to $p_{\theta_{k-1}}\left(\mathsf{x}|\mathsf{y}\right)p_{\theta^*}(\mathsf{y})$ and $\Psi\left(\mathsf{X},\mathsf{Y}\right)=\left[\psi_1\left(\mathsf{X},\mathsf{Y}\right),\psi_2\left(\mathsf{X},\mathsf{Y}\right),\psi_3\left(\mathsf{X},\mathsf{Y}\right),\psi_4\left(\mathsf{X},\mathsf{Y}\right)\right]^T$ with $\psi_1\left(\mathsf{X},\mathsf{Y}\right)=X_1^2+X_L^2$, $\psi_2\left(\mathsf{X},\mathsf{Y}\right)=\sum_{i=2}^{L-1}X_i^2$, $\psi_3\left(\mathsf{X},\mathsf{Y}\right)=\sum_{i=2}^{L}X_iX_{i-1}$, $\psi_4\left(\mathsf{X},\mathsf{Y}\right)=\sum_{i=1}^{L}Y_i^2\exp\left(-X_i\right)$. Given $\Phi\left(\theta_{k-1},\theta^*\right)$, it is possible to maximize $Q\left(\theta,\theta_{k-1}\right)$ analytically when $L\geq 2$ and find $\theta_k=\Lambda(\Phi\left(\theta_{k-1},\theta^*\right))$ where Λ is not given here for sake of brevity.

In practice, Q cannot be computed as the expectations appearing in the expression for $\Phi\left(\theta_k,\theta^*\right)$ are with respect to a measure dependent on the unknown parameter value θ^*. However, this ideal batch algorithm can be approximated using the following on-line scheme. Indeed, thanks to the ergodicity and stationarity assumptions, the observations $\{\mathsf{Y}_k\}$ provide us with samples from $p_{\theta^*}(\mathsf{y})$ which can be used for the purpose of Monte Carlo integration. More precisely we recursively approximate the sufficient statistics $\Phi\left(\theta_k,\theta^*\right)$ with the following update, given here at time k,

$$\hat{\Phi}_k=\left(1-\gamma_k\right)\hat{\Phi}_{k-1}+\gamma_k\mathbb{E}_{\theta_{k-1}}\left(\left.\Psi\left(\mathsf{X},\mathsf{Y}_k\right)\right|\mathsf{Y}_k\right)\ , \tag{10}$$

where $\mathbb{E}_{\theta_{k-1}}\left(\cdot|\mathsf{Y}_k\right)$ denotes the expectation with respect to $p_{\theta_{k-1}}\left(\mathsf{x}|\mathsf{Y}_k\right)$. We then substitute $\hat{\Phi}_k$ for $\Phi\left(\theta_k,\theta^*\right)$ and obtain $\theta_k=\Lambda(\hat{\Phi}_k)$. If θ_k was constant and $\gamma_k=k^{-1}$ then $\hat{\Phi}_k$ would simply compute the arithmetic average of $\{\mathbb{E}_{\theta_{k-1}}(\left.\Psi\left(\mathsf{X},\mathsf{Y}_k\right)\right|\mathsf{Y}_k)\}$, and converge towards $\Phi\left(\theta_k,\theta^*\right)$ by ergodicity. In fact, convergence is in general ensured for any non-increasing positive stepsize sequence $\{\gamma_k\}$ such that $\sum\gamma_k=\infty$ and $\sum\gamma_k^2<\infty$; we can select $\gamma_k=C.k^{-\alpha}$ where $C>0$ and $\alpha\in\left(\frac{1}{2},1\right]$.

To summarize, the vector of sufficient statistics $\hat{\Phi}_{-1}$ is arbitrarily initialized and the on-line EM algorithm proceeds as follows for the data slice indexed by $k\geq 0$.

E-step: $\hat{\Phi}_k=(1-\gamma_k)\hat{\Phi}_{k-1}+\gamma_k\mathbb{E}_{\theta_{k-1}}\left(\left.\Psi\left(\mathsf{X},\mathsf{Y}_k\right)\right|\mathsf{Y}_k\right)$.

M-step: $\theta_k=\Lambda(\hat{\Phi}_k)$.

In scenarios where $\mathbb{E}_{\theta_k}\left(\left.\Psi\left(\mathsf{X},\mathsf{Y}_k\right)\right|\mathsf{Y}_k\right)$ does not admit an analytical expression, a further Monte Carlo approximation can be used. Assume that a good approximation $q_\theta\left(\cdot|\mathsf{Y}_k\right)$ of $p_\theta\left(\cdot|\mathsf{Y}_k\right)$ is available, and that it is easy to sample from $q_\theta\left(\cdot|\mathsf{Y}_k\right)$. In this case the algorithm presented above can be altered as follows.

E-step $\mathsf{X}_k^{(i)}\sim q_{\theta_{k-1}}\left(\cdot|\mathsf{Y}_k\right)$ for $i=1,\ldots,N$

$\hat{\Phi}_k=(1-\gamma_k)\hat{\Phi}_{k-1}+\gamma_k\sum_{i=1}^N W_k^{(i)}\Psi(\mathsf{X}_k^{(i)},\mathsf{Y}_k)$ where $W_k^{(i)}\propto\frac{p_{\theta_{k-1}}(\mathsf{X}_k^{(i)}|\mathsf{Y}_k)}{q_{\theta_{k-1}}(\mathsf{X}_k^{(i)}|\mathsf{Y}_k)}$, and $\sum_{i=1}^N W_k^{(i)}=1$.

M-step: $\theta_k=\Lambda(\hat{\Phi}_k)$.

As N increases the importance sampling approximation converges towards the true expectation. Moreover the bias is of order N^{-1}. Note that if it is possible to sample from $p_{\theta_{k-1}}\left(\mathsf{x}|\mathsf{Y}_k\right)$ exactly then it is not necessary to have a large number N of samples and a single sample might even be sufficient. Indeed it is only necessary to have an unbiased estimate of $\mathbb{E}_{\theta_{k-1}}\left(\left.\Psi\left(\mathsf{X},\mathsf{Y}_k\right)\right|\mathsf{Y}_k\right)$.

Observe also that it is alternatively possible to use SMC techniques to approximate this expectation or to sample approximately from $p_\theta\left(\mathsf{x}|\mathsf{Y}_k\right)$. We stress that in this context, as SMC is used to sample from a distribution of fixed dimension L, there will be no path degeneracy problem.

C. *On-line EM Algorithm: Unknown Invariant Distribution*

In the previous section we required not only the existence of π_θ but also its analytical expression. This can be a restriction in some situations since whereas ergodicity can be established for many Markov processes of interest, and the existence of π_θ proved, closed form expression for this distribution is rarely available due to algebraic intractability. However in many cases, it is easy to sample realizations of the process $\{X_n,Y_n\}$ for a fixed value of θ, especially in situations where the process is defined recursively as

$$X_{n+1}=\varphi_\theta\left(X_n,V_{n+1}\right),\ Y_n=\gamma_\theta\left(X_n,W_n\right), \tag{11}$$

where for any $\theta \in \Theta$, $\varphi_\theta : \mathsf{X} \times \mathsf{V} \to \mathsf{X}$, $\gamma_\theta : \mathsf{X} \times \mathsf{W} \to \mathsf{X}$ for some spaces V and W, with $\{V_n\} \subset \mathsf{V}^{\mathbb{N}}$ and $\{W_n\} \subset \mathsf{W}^{\mathbb{N}}$ being *i.i.d.* sequences from standard distributions not dependent on θ. Here we will consider the situation where we can sample realizations of the process $\{X_n, Y_n\}$ for a fixed value of θ to develop an algorithm based on the *indirect inference* principle; *e.g.* [8]. The main idea of indirect inference consists of the following three key steps. First a proxy model parameterized by $\xi \in \Xi$ is fitted to the observed data $\{Y_n^*\}$, generated with $\theta = \theta^*$. We will hereafter denote $\xi^* = \xi(\theta^*)$ the corresponding estimate. The proxy model is generally chosen so that inference is easy, but sufficiently close to the original model in order to capture its full complexity. Second, given a parameter estimate $\hat{\theta}$ of the true parameter θ^* one can simulate artificial data $\{\hat{Y}_n\}$ using (11) with $\theta = \hat{\theta}$, and the aforementioned proxy model is also fitted to $\{\hat{Y}_n\}$ to produce an estimate $\hat{\xi} = \xi(\hat{\theta})$. Third, the parameter estimate $\hat{\theta}$ is updated to decrease a distance measure between $\hat{\xi}$ and ξ^*. These three steps are repeated until $\hat{\xi} = \xi^*$. Many criteria are possible for matching, and we chose here the commonly used criterion

$$J(\theta) := (\xi - \xi^*)^{\mathsf{T}} \ \Sigma \ (\xi - \xi^*) \ , \tag{12}$$

where $\xi = \xi(\theta)$ and Σ is a positive definite matrix; see [8] and the references therein for a discussion of the validity of the approach. There is, in general, no analytical expression for the minimizer of J, and we will resort here to a steepest descent algorithm, which requires the computation of the gradient

$$\frac{1}{2}\nabla J(\theta) = \nabla\xi \ \Sigma \ (\xi - \xi^*) \ . \tag{13}$$

Note that here ∇ denotes the derivative with respect to θ and that ∇_z will denote the gradient with respect to any other variable z. Note also that for an n_z-dimensional variable $z = [z_1 \dots z_{n_z}]$ and an n_h-dimensional function $h = [h_1 \dots h_{n_h}]^{\mathsf{T}}$, we will use the notation $\nabla_z h$ to denote the $n_z \times n_h$ matrix with elements$[\nabla_z h]_{i,j} = \partial h_j / \partial z_i$. This methodology is very general and good performance is in general obtained if the auxiliary model is "close" to the true model. We suggest here a proxy model $\bar{p}_\xi(\mathsf{x}_k, \mathsf{y}_k)$ which differs from the true model (8) only in that α_ξ replaces the invariant distribution π_θ. Here ξ and θ coincide, *i.e.* $\xi \in \Xi = \Theta$, if $\alpha_\xi = \alpha$. Following the developments of Section III-B, we introduce the following cost function for any $(\xi, \theta) \in \Xi \times \Theta$

$$\tilde{l}_\theta(\xi) := \int_{\mathsf{Y}^L} \log(\bar{p}_\xi(\mathsf{y}))\, p_\theta(\mathsf{y})\, d\mathsf{y} \ ,$$

and for $\hat{\theta}, \theta^* \in \Theta$ define the "pseudo-estimates" $\hat{\xi}, \xi^*$ as

$$\hat{\xi} := \underset{\xi \in \Xi}{\arg\max}\ \tilde{l}_{\hat{\theta}}(\xi) \ \text{ and } \xi^* := \underset{\xi \in \Xi}{\arg\max}\ \tilde{l}_{\theta^*}(\xi) \ .$$

In most cases analytical expressions for $\hat{\xi}, \xi^*$ and $\nabla\hat{\xi} = \nabla\xi|_{\theta=\hat{\theta}}$ are not available, and we resort to iterative methods. Following the developments of Section III-B, we again suggest the use of an on-line EM algorithm in order to construct sequences $\{\hat{\xi}_k\}$ and $\{\xi_k^*\}$ which converge to estimates of $\hat{\xi}$ and ξ^*. Typically, for a given $\theta \in \Theta$, $\{\xi_k\}$ is defined recursively as $\xi_k = \Lambda(\Psi(\xi_{k-1}, \theta))$ where $\Psi(\xi, \theta)$ is a vector of sufficient statistics and Λ a deterministic mapping.

In most cases of interest this "ideal" algorithm cannot be implemented as the expectations do not admit closed-form expressions and, in particular, θ^* is itself unknown. However, we can again resort to on-line Monte Carlo approximations. For any θ, integration with respect to $p_\theta(\mathsf{y})$ can be performed by generating artificial data $\{Y_k\}$ from (11) and using the recursion

$$\Phi_k = (1 - \gamma_k)\, \Phi_{k-1} + \gamma_k \int_{\mathsf{X}^L} \Psi(\mathsf{x}, \mathsf{Y}_k)\, \bar{p}_{\xi_{k-1}}(\mathsf{x} | \mathsf{Y}_k)\, d\mathsf{x} \ , \tag{14}$$

for a stepsize sequence $\{\gamma_k\} \subset (0,1)^{\mathbb{N}}$. The sequence $\{\xi_k\}$ is constructed by letting $\xi_k = \Lambda(\Phi_k)$ for any $k \geq 1$. We refer to this algorithm as the on-line EM algorithm. Similarly integration with respect to $p_{\theta^*}(\mathsf{y})$ is straightforward, since here $\{Y_k^*\}$ provides us directly with samples which can be used for Monte Carlo integration and fed into a recursion identical to (14) to produce $\{\Phi_k^*\}$. If integration with respect to $\bar{p}_\xi(\mathsf{x}|\mathsf{y})$, denoted hereafter $\mathbb{E}_\xi(\cdot|\mathsf{y})$, is not possible analytically, it can be performed using importance sampling or more generally any other Monte Carlo technique, similarly to that in Section III-B. We now focus on a technique to construct sequences $\{\nabla\hat{\xi}_k\}$ which converges to an estimate of $\nabla\hat{\xi}$, as this quantity is required in order to minimize $J(\theta)$.

To this end, for any $k \geq 1$ we consider the gradient with respect to θ of ξ_k produced by the on-line EM algorithm for a set of observations $\{Y_n\}$ generated by (11) with θ. First it is worth recalling that ξ_k is obtained at iteration k by a deterministic transformation of the estimated sufficient statistics Φ_k of $\Phi(\xi_{k-1}, \theta)$, $\xi_k = \Lambda(\Phi_k)$. As a consequence the derivative with respect to θ is of the form

$$\nabla\xi_k = \nabla\Phi_k . \nabla_{\Phi_k} \Lambda(\Phi_k) \ . \tag{15}$$

A sequence $\{\nabla\Phi_k\}$ of gradients of the sufficient statistics can be recursively constructed by differentiating (14), leading to

$$\nabla\Phi_k = (1 - \gamma_k)\, \nabla\Phi_{k-1} + \gamma_k \nabla \int_{\mathsf{X}^L} \Psi(\mathsf{x}_k, \mathsf{Y}_k)\, \bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k)\, d\mathsf{x}_k \ , \tag{16}$$

where $\{Y_k\}$ is generated using θ. Under regularity assumptions, it follows that $\nabla \int_{\mathsf{X}^L} \Psi(\mathsf{x}_k, \mathsf{Y}_k)\, \bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k)\, d\mathsf{x}$ is equal to

$$\begin{aligned}&\int_{\mathsf{X}^L} \nabla\mathsf{Y}_k . \nabla_{\mathsf{Y}} \Psi(\mathsf{x}_k, \mathsf{Y}_k)\, \bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k)\, d\mathsf{x}_k \\ &+ \int_{\mathsf{X}^L} \nabla\bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k) . \Psi(\mathsf{x}_k, \mathsf{Y}_k)^{\mathsf{T}}\, d\mathsf{x}_k \ ,\end{aligned}$$

where $\nabla \log \bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k)$ is given by

$$\begin{aligned}&\nabla \log \bar{p}_{\xi_{k-1}}(\mathsf{x}_k, \mathsf{Y}_k) - \\ &\int_{\mathsf{X}^L} \nabla \log \bar{p}_{\xi_{k-1}}(\mathsf{x}_k, \mathsf{Y}_k)\, \bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k)\, d\mathsf{x}_k \bar{p}_{\xi_{k-1}}(\mathsf{x}_k | \mathsf{Y}_k)\end{aligned} \tag{18}$$

providing us with an expression for (17) in terms of expectations $\mathbb{E}_{\xi_{k-1}}(\cdot|\mathsf{Y}_k)$. Finally we have $\nabla \log \bar{p}_{\xi_{k-1}}(\mathsf{x}_k, \mathsf{Y}_k)$ equal to

$$\sum_{n=kL+2}^{(k+1)L} \nabla \log f_{\xi_{k-1}}(x_n | x_{n-1}) + \sum_{n=kL+1}^{(k+1)L} \nabla \log g_{\xi_{k-1}}(Y_n | x_n) \ ,$$

and $\nabla \log f_{\xi_{k-1}}(x_n | x_{n-1}) = \nabla\xi_{k-1} . \nabla_{\xi_{k-1}} \log f_{\xi_{k-1}}(x_n | x_{n-1})$,

$$\begin{aligned}\nabla \log g_{\xi_{k-1}}(Y_n | x_n) = \nabla Y_n . \nabla_y \log g_{\xi_{k-1}}(Y_n | x_n) \\ + \nabla\xi_{k-1} . \nabla_{\xi_{k-1}} \log g_{\xi_{k-1}}(Y_n | x_n) \ .\end{aligned}$$

The sequence $\{\nabla Y_n\}$ required for $\{\nabla\mathsf{Y}_k\}$ corresponds to a path derivative [9]. Assuming that we can sample $\{V_n\}$ and $\{W_n\}$ exactly (recall these are independent of θ) and that the functions φ_θ and γ_θ are differentiable with respect to their first argument, the sequence can be recursively computed as follows $\nabla X_0 = 0$, $\nabla Y_0 = 0$ and for $n > 0$

$$\nabla X_{n+1} = \nabla Y_n . \nabla_x \varphi_\theta(X_n, V_{n+1}) + \nabla\varphi_\theta(X_n, V_{n+1}) \tag{19}$$

$$\nabla Y_n = \nabla X_n . \nabla_x \gamma_\theta(X_n, W_n) + \nabla\gamma_\theta(X_n, W_n) \ . \tag{20}$$

Assuming that all the above expectations with respect to $\bar{p}_{\xi_{k-1}}(\mathsf{x}_k|\,\mathsf{Y}_k)$ can be calculated analytically, (14) for $\theta = \hat{\theta}, \theta^*$ will allow us to compute sequences $\{\hat{\xi}_k\}$ and $\{\xi_k^*\}$ that will converge to estimates of $\hat{\xi}$ and ξ^* and (20) for $\theta = \hat{\theta}$ will allow us to compute a sequence $\{\nabla\hat{\xi}_k\}$ that will converge to $\nabla\hat{\xi}$, therefore allowing us to compute an estimate of $\frac{1}{2}\nabla J|_{\theta=\hat{\theta}}$.

One can therefore, in theory, construct a sequence $\{\hat{\theta}_l\}$ that will converge to the set of stationary points of $\frac{1}{2}\nabla J$ using the recursion

$$\hat{\theta}_l = \hat{\theta}_{l-1} - \gamma_l \ \nabla\hat{\xi}^l \ \Sigma \ (\hat{\xi}^l - \xi^*) \ , \qquad (21)$$

where for $l \geq 1$ $\hat{\xi}^l := \lim_{k\to\infty} \hat{\xi}_k^l$, $\xi^* := \lim_{k\to\infty} \hat{\xi}_k^*$ and $\nabla\hat{\xi}^l := \lim_{k\to\infty} \nabla\hat{\xi}_k^l$ with $\{\hat{\xi}_k^l\}$ estimated for $\theta = \hat{\theta}_{l-1}, \theta^*$ and $\{\nabla\hat{\xi}_k^l\}$ estimated for $\theta = \hat{\theta}_{l-1}$. This is clearly impossible in practice. A first step towards preserving the on-line nature of our algorithm would be to redefine $\hat{\xi}^l := \hat{\xi}_{k_l}^l$, $\xi^* := \hat{\xi}_{k_l}^*$ and $\nabla\hat{\xi}^l := \nabla\hat{\xi}_{k_l}^l$ for a sequence of integers $\{k_l\}$ such that $\lim_{l\to\infty} k_l = \infty$. This is however still unsatisfactory since the computational cost per iteration (21) grows to infinity with l and the algorithm does not "recycle" estimates $\hat{\xi}^l$ from one iteration to the next.

An elegant way around these two problems consists of using a two-time scale scheme, where the auxiliary model if fitted on a fast timescale whereas the estimate of the true parameter is updated on a slow timescale. This scheme requires the choice of two non-decreasing positive stepsize sequences $\{\gamma_k\}$, $\{\beta_k\}$ such that

$$\sum_{k\geq 1} \gamma_k = \sum_{k\geq 1} \beta_k = \infty \ , \quad \sum_{k\geq 1} \gamma_k^2 < \infty \ , \quad \sum_{k\geq 1} \beta_k^2 < \infty \ ,$$

and for some $\delta > 0$, $\sum_{k\geq 1} \beta_k^\delta \gamma_k^{-\delta} < \infty$. A typical choice is $\gamma_k = C_1 k^{-\nu}$, $\beta_k = C_2 k^{-\zeta}$ for $1 > \zeta > \nu > 0.5$. The algorithm is initialized with arbitrary values $\hat{X}_0, \hat{Y}_0, \nabla\hat{X}_0, \nabla\hat{Y}_0, \Phi_{-1}^*$ and $\hat{\Phi}_{-1}$ and then proceeds as follows for $k \geq 0$.

EM for the true data/proxy model

$$\Phi_k^* = (1-\gamma_k)\Phi_{k-1}^* + \gamma_k \mathbb{E}_{\xi_{k-1}^*}\left(\Psi(\mathsf{X}, \mathsf{Y}_k^*)|\,\mathsf{Y}_k^*\right) \ , \xi_k^* = \Lambda(\Phi_k^*) \ .$$

Sampling of artificial data For $n = kL+1, kL+2, \ldots, (k+1)\,L$, sample $\hat{V}_n$ and $\hat{W}_n$ then set

$$\hat{X}_n = \varphi_{\hat{\theta}_{k-1}}(\hat{X}_{n-1}, \hat{V}_n) \ , \hat{Y}_n = \gamma_{\hat{\theta}_{k-1}}(\hat{X}_n, \hat{W}_n) \ .$$

EM for the artificial data/proxy model

$$\hat{\Phi}_k = (1-\gamma_k)\hat{\Phi}_{k-1} + \gamma_k \mathbb{E}_{\hat{\xi}_{k-1}}(\Psi(\mathsf{X}, \hat{\mathsf{Y}}_k)|\hat{\mathsf{Y}}_k) \ , \hat{\xi}_k = \Lambda(\hat{\Phi}_k) \ .$$

Model Matching, compute $\nabla\hat{\xi}_k$ using (15) to (20) and

$$\hat{\theta}_k = \hat{\theta}_{k-1} - \beta_k \ \nabla\hat{\xi}_k \ \Sigma \ (\hat{\xi}_k - \xi_k^*) \ ,$$

The expectations with respect to $\bar{p}_{\xi_{k-1}^*}(\mathsf{x}|\,\mathsf{Y}_k^*)$ and $\bar{p}_{\hat{\xi}_{k-1}}(\mathsf{x}|\,\hat{\mathsf{Y}}_k))$ appearing in $\mathbb{E}_{\xi_{k-1}^*}(\Psi(\mathsf{X}, \mathsf{Y}_k^*)|\,\mathsf{Y}_k^*)$ and $\mathbb{E}_{\hat{\xi}_{k-1}}(\Psi(\mathsf{X}, \hat{\mathsf{Y}}_k)|\hat{\mathsf{Y}}_k)$, respectively, cannot typically be performed in closed-form. Monte Carlo methods similar to those described in the previous section can be used. A Monte Carlo approximation of $\bar{p}_{\hat{\xi}_{k-1}}(\mathsf{x}|\,\hat{\mathsf{Y}}_k)$ can also be used to estimate $\nabla\hat{\xi}_k$ in (18).

IV. Application

We have applied our algorithms to the stochastic volatility model given in Example 2. Simulation results for the algorithm based on indirect inference are presented in the extended version of this article. The sampling distribution q_θ was chosen to be a Gaussian approximation of p_θ as described in [10]. In Figure 3, we present a simulation for the case where $L = 10$, $N = 100$ $\phi = 0.8$, $\sigma_v^2 = 0.1$ and $\beta^2 = 1$, for a number of observations $T = 250,000$ and $\gamma_k = 1/k^{1/2}$. We also present results corresponding to the on-line EM algorithm and a modified version of it using the Polyak-Ruppert averaging procedure. In this case, this proves very useful for the parameter β. The algorithm converges to the true parameter. We also display in Figure 3 the Kullback-Leibler divergence between the estimated parameters and the true parameters.

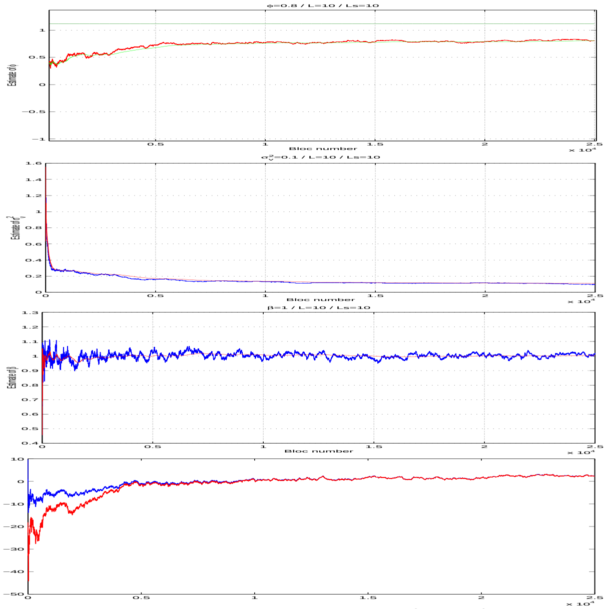

Fig. 3. From top to bottom: Convergence of ϕ, σ_v^2 and β^2 and comparison (up to additive constants) of the KL divergence under the estimated parameters (red) and the true parameters (blue).

References

[1] Andrieu, C., De Freitas, J.F.G. and Doucet, A. (1999). Sequential MCMC for Bayesian model selection. *Proceedings IEEE Work. HOS*, 130-134.

[2] Andrieu, C., Doucet, A., Singh, S.S. and Tadić, V.B. (2004) Particle Methods for Change Detection, Identification and Control. *Proceedings of the IEEE*, 92, 423-438.

[3] Benveniste, A., Métivier, M. and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximation*. New York: Springer-Verlag.

[4] Del Moral, P. (2004) *Feynman-Kac formulae. Genealogical and interacting particle approximations.* Springer New York, Series: Probability and Applications.

[5] Doucet, A., de Freitas, J.F.G. and Gordon N.J. (eds.) (2001). *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag.

[6] Fearnhead, P. (2002) MCMC, sufficient statistics and particle filter. *J. Comp. Graph. Stat.*, 11, 848-862.

[7] Gilks, W.R. and Berzuini, C. (2001) Following a moving target - Monte Carlo inference for dynamic Bayesian models, *J.R. Statist. Soc.* B, 63, 127-146.

[8] Heggland, K. and Frigessi, A. (2004). Estimating functions in indirect inference. *J.R. Statist. Soc.* B, 66, 447-462.

[9] Pflug, G.C. (1996). *Optimization of stochastic models: The interface between simulation and optimization*. Boston: Kluwer.

[10] Pitt, M.K. and Shephard, N. (1999). Filtering via simulation: auxiliary particle filter. *J. Am. Statist. Ass.*, 94, 590-599.

[11] Poyiadjis, G., Doucet, A. and Singh S.S. (2005) Particle method for optimal filter derivative: Application to parameter estimation. *Proceedings IEEE ICASSP*, 5, 925-928.

[12] Rydén T. (1997) On recursive estimation for hidden Markov models. *Stochastic Processes and their Applications*, 66, 79-96.

[13] Storvik G. (2002) Particle filters in state space models with the presence of unknown static parameters. *IEEE. Trans. Signal Processing*, 50, 281–289.